

Multiframe Motion Coupling via Infimal Convolution Regularization for Video Super Resolution

Hendrik Dirks*, Jonas Geiping†, Daniel Cremers‡, Michael Moeller§

November 24, 2016

Abstract

The idea of video super resolution is to use the temporal information of seeing a scene from many slightly different viewpoints in the successive frames of a video to enhance the overall resolution and quality of each frame. Classical energy minimization approaches first establish a correspondence of the current video frame to several of its neighbors and then use this temporal information to enhance it. In this paper we propose the first variational super resolution approach that computes several super resolved frames in one joint optimization procedure by incorporating motion information between the high resolution image frames themselves. As a consequence, the number of motion estimation problems grows linearly in the number of frames, opposed to a quadratic growth of classical methods.

In addition, we use infimal convolution regularization to automatically determine the reliability of the motion information and reweight the regularization locally. We demonstrate that our approach yields state-of-the-art results and even is competitive with learning based approaches that require a significant amount of training data.

1 Introduction

The technique of video super resolution combines the spatial information from several low resolution frames of the same scene to produce a high resolution video.

A classical way of solving the super resolution problem is to estimate the motion from the current frame to m neighboring frames, model the data formation process via a warping, blur, and downsampling, and use a suitable regularization to suppress possible artifacts arising from the ill-posedness of the underlying problem. The final goal is to produce an enhanced, visually pleasing high resolution video in a reasonable runtime.

Note that the super resolution of n video frames using the classical approach described above solves n separate problems - one for each frame. Each of these problems requires

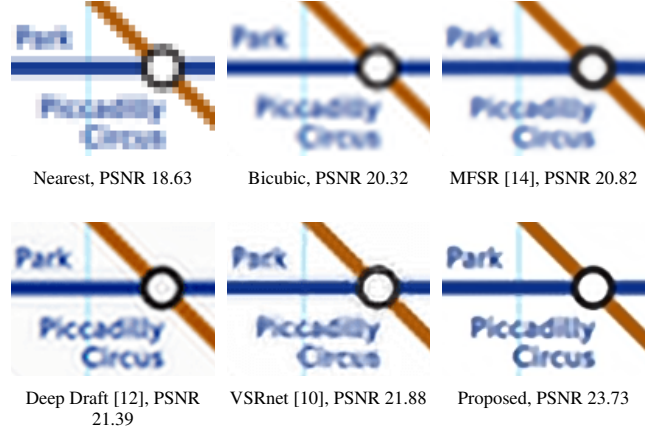


Figure 1: Results for super resolving a set of 13 images of a London tube map by a factor of 4. While the text is not readable in the low resolution image, super resolution techniques can exploit the temporal information to improve the visual quality significantly. Due to the idea of jointly super resolving multiple frames, our approach behaves superior to the competing variational approach [14]. While approaches based on learning with massive amounts of data are competitive to our approach on real world scenarios (see section 4), our method outperforms these methods significantly on examples which likely differ from the usual training data.

m motion estimations. To introduce the maximum amount of temporal information one has to choose $m = n - 1$ and the total number of flow estimations $n(n - 1)$ increases quadratically with the number of frames. Moreover, due to the strategy of super resolving each frame separately, temporal consistency cannot be enforced explicitly. The latter, however, is a key feature for a video to be visually pleasing: Even if a method generates a sequence of high quality high resolution frames, temporal inconsistencies will be visible as a disturbing flickering.

In addition, choosing the right strength of the regularization is a delicate issue. While a small regularization allows significant improvements in areas where the motion estimation is precise, it can lead to heavy oscillations and ringing artifacts in areas of quick motion and occlusions. A large regularization on the other hand avoids these artifacts but quickly

*University of Münster, hendrik.dirks@wwu.de

†University of Siegen, jonas.geiping@uni-siegen.de

‡Technical University of Munich, cremers@tum.de

§University of Siegen, michael.moeller@uni-siegen.de

oversmooths the image and hence also suppresses the desirable super resolution effect.

1.1 Contribution

We propose a method that jointly solves for the whole super resolved video and couples neighboring high resolution frames directly. Such an approach tackles two of the drawbacks mentioned above: Because only neighboring frames are coupled, the number of required motion estimations grows linearly with the number of frames. By introducing this coupling on the unknown high resolution images directly, the appearance of each frame still influences all other frames and information can be exchanged over the entire sequence. In particular, the problem does not decouple and therefore enforces temporal consistency explicitly.

Furthermore, we tackle the problem of choosing the right strength of spatial regularity by proposing to use the *infimal convolution* between a strong spatial and a strong temporal regularization term. The latter allows our framework to automatically select the right type of regularization locally in a single convex optimization approach that can be minimized globally. The proposed infimal convolution regularization provides robustness to inconsistencies in the motion fields and at the same time avoids overregularization in areas that profit from the additional temporal information.

As illustrated in Figure 1 our approach yields state-of-the-art results which are even competitive with learning based approaches that require significant amounts of training data. While Figure 1 is a synthetic test consisting of planar motion only, we demonstrate the performance of the proposed approach on several real world videos in Section 4. In summary:

- To the best of our knowledge we propose the first variational super resolution method that uses motion information to couple the high resolution frames directly.
- We use an infimal convolution regularization between terms that focus on spatial and temporal regularity respectively yielding a method that is robust to errors in the estimated motion fields.
- The number of required optical flow estimations grows only linearly with the number of frames.
- We jointly optimize for all frames of the video simultaneously and therefore enforce temporal consistency directly.
- The method generates state-of-the-art results on challenging datasets. Without requiring any training data it performs on par with recent deep learning approaches.

1.2 State of the Art

The literature on super resolution techniques is vast and it goes beyond the scope of the paper to present a complete

overview. An extensive survey about super resolution techniques published before 2012 can be found in [17]. We will focus on recalling some recent approaches based on energy minimization and deep learning techniques.

Variational Video Reconstruction

Most variational approaches for super resolution consist of three terms; a *data fidelity term* that relates each high resolution frame to its low resolution counterpart, one *temporal consistency term* that exploits the additional information hidden in neighboring frames via a motion-based coupling, and a *regularization term* to tackle the ill-posedness of the underlying problem and enforce spatial regularity.

We refer the reader to [8] for a comprehensive description of the data formation process and the resulting energies. An example of a variational method based on the three terms discussed above is [24], in which the authors propose to determine a high resolution version of the i -th frame by minimizing

$$\min_{u^i} \|D(b * u^i) - f^i\|_{H^{\epsilon_d}} + \lambda \|\nabla u^i\|_{H^{\epsilon_r}} + \sum_{j \neq i} \|D(b * W^{j,i} u^i) - f^j\|_{H^{\epsilon_d}}, \quad (1)$$

where $\|\cdot\|_{H^{\epsilon_d}}$ denotes the Huber loss, D a downsampling operator, b a blur kernel, λ a regularization parameter, and $W^{j,i}$ a warping operator that compensates the motion from the j -th to the i -th frame and is computed by an optical flow estimation in a first processing step. The temporal consistency term is based on $D(b * W^{j,i} u^i) - f^j$ and hence compares each frame to multiple low resolution frames. It is important to note that repeating the minimization (1) for super resolving a video means that the high resolution frames are not coupled directly and one requires $n(n-1)$ optical flow estimations for n being the total number of frames.

In [13] Liu and Sun proposed to incorporate different (global) weights $\theta_{j,i}$ for each of the temporal consistency terms, and additionally estimate the blur kernel b , the warping operators $W^{j,i}$ by applying alternating minimization.

In [14] Ma et al. extended the work [13] for the case of some of the low resolution frames being particularly blurry. Instead of the global weight parameters they introduce a localized version that tries to determine the value of the temporal regularization on a pixel level in the framework of an expectation maximization (EM) algorithm. Moreover, a non-convex truncated quadratic penalty on the gradient of u^i was used and minimized using reweighting techniques.

Similar to (1) the energies proposed in [13, 14] do not enforce regularity between the high resolution frames u^i directly and require quadratically many motion estimations.

The framework by Shechtman et al. [21] incorporates a motion blur in time and space together with L^2 -regularization. The problem is solved in the full space-time domain but does not incorporate any motion information.

In a recent work [4] on time continuous variational models, the authors proposed to use an optical flow penalty $\|\nabla u \cdot v + u_t\|_1$ as a temporal regularization for reconstructing a given image sequence u and calculating flow fields v between subsequent images at the same time. While the optical flow term is exact in the temporally continuous setting, it would require small motions of less than one pixel to be a good approximation in a temporally discrete video consisting of separate frames. Although such a small-motion-assumption does not hold for the super resolution problem considered here, the work [4] still motivates our temporal consistency term that directly couples the (unknown) high resolution video frames.

Learning based approaches

With the recent breakthroughs of deep learning and convolutional neural networks, researchers have promoted learning-based methods for video super resolution [10, 12, 30, 11]. These seem to be the currently leading techniques for video super resolution.

The focus of [30] is the development of a real-time capable super resolution technique, such that we will concentrate our comparison to [12], [10], and [11], which focus on high image quality rather than computational efficiency.

Note that [12] and [10] work with motion correction and require optical flow estimations. Similar to the classical variational techniques they register multiple neighboring frames to the current frame and hence also require quadratically many flow estimations.

The very deep convolutional network [11] proposed by Kim et al. is a conceptually different approach that does not use any temporal information, but solely relies on the training data.

In the current deep learning hype, one should keep in mind, however, that learning-based and non-learning based approaches address two very different problems: Whereas the variational approaches aim at recovering a super resolution video from the low-resolution video with a minimal set of regularity assumptions, learning-based approaches take as input the low-resolution video and millions of training examples. Although the proposed variational approach does not require any training data and thus a comparison of these techniques does not appear fair, we include state-of-the-art deep learning techniques in the evaluations for completeness. In fact, we will see that the proposed method performs quite on par with learning-based approaches.

2 Proposed Method

For a sequence $f = f^1, \dots, f^n$ of low-resolution input images we propose a multi-frame super resolution model based on motion coupling between subsequent frames. Opposed to any of the variational approaches summarized in the previous section, the energy we propose directly couples all (un-

known) high resolution frames. Our method jointly computes the super resolved versions of n video frames at once via the following minimization problem,

$$\min_u \sum_{i=1}^n \|D(b * u^i) - f^i\|_1 + \alpha(R_{\text{temp}} \square R_{\text{spat}})(u). \quad (2)$$

The first term is a standard data fidelity term similar to (1). The key novelty of our approach is twofold and lies in the way we incorporate and utilize the motion information as well as the way we combine the temporal information with a spatial regularity assumption. The latter is an extension of a spatio-temporal infimal convolution technique proposed by Holler and Kunisch in [9] for video denoising and decompression.

2.1 Spatio-Temporal Infimal Convolution

The second term in (2) denotes the *infimal convolution* [5] between a term R_{temp} , which is mostly focused on introducing temporal information, and a term R_{spat} , which is mostly focused on enforcing spatial regularity on u . The infimal convolution between the two terms is defined as

$$(R_{\text{temp}} \square R_{\text{spat}})(u) := \inf_{u=w+z} R_{\text{temp}}(w) + R_{\text{spat}}(z). \quad (3)$$

It can be understood as a logical OR connection and allows to optimally divide the input u into two parts, one of which is preferable in terms of the costs R_{temp} and the other one in terms of the costs R_{spat} . The respective costs are defined as

$$R_{\text{spat}}(u) = \sum_{i=1}^n \left\| \sqrt{(u_x^i)^2 + (u_y^i)^2 + (\kappa W(u^i, u^{i+1}))^2} \right\|_1,$$

$$R_{\text{temp}}(u) = \sum_{i=1}^n \left\| \sqrt{(\kappa u_x^i)^2 + (\kappa u_y^i)^2 + (W(u^i, u^{i+1}))^2} \right\|_1,$$

for $\kappa < 1$, where the subscripts x and y denote the x - and y -derivatives, and W denotes the photoconsistency

$$W(u^i, u^{i+1})(x) = \begin{cases} \frac{u^i(x) - u^{i+1}(x + v^i(x))}{h} & \text{if } i \text{ even, } i < n, \\ \frac{u^{i+1}(x) - u^i(x + v^i(x))}{h} & \text{if } i \text{ odd, } i < n, \\ 0 & \text{if } i = n. \end{cases}$$

The idea for using such an infimal convolution approach originates from [9] in which the authors used a similar term with a time derivative instead of the operator W for video denoising and decompression. The infimal convolution automatically selects a regularization focusing either on space or time at each point. At points in the image where the warp energy $W(u^i, u^{i+1})$ is high, our approach automatically uses strong total variation (TV) regularization. In this sense it is an elegant way of replacing the EM-based local parameter estimation from [14] by a joint and fully automatic regularization method with similar effects: It can handle inconsistencies in the motion field v by deciding to determine such locations by R_{spat} . On the other hand introducing strong spatial regularity

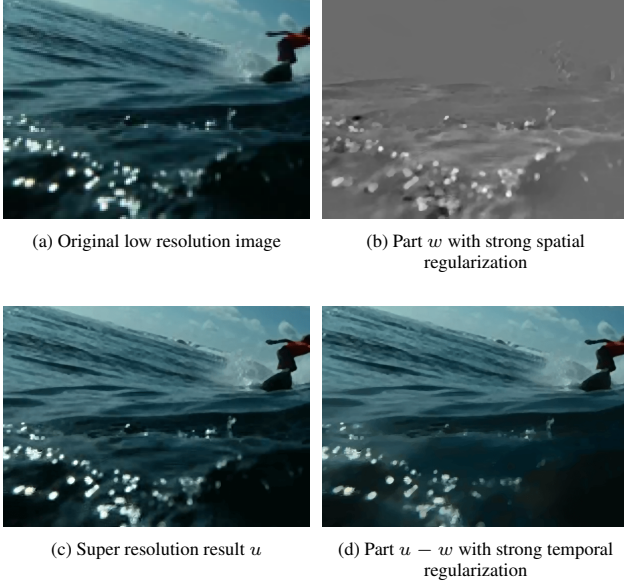


Figure 2: Illustrating the behavior of infimal convolution regularization: The super resolution result u of the low resolution input data from (a) is given in Subfigure (c). Subfigures (b) and (c) illustrate the division of u into the two parts w and $z = u - w$ determined by the infimal convolution regularization (3). Our method automatically imposes more spatial regularity in the (random) reflections of the wave front for which the optical flow cannot provide reasonable motion information. Background and sky are far enough away to exhibit rather slow motion such that they are regularized temporally. Note that w has been rescaled (and kept as a grayscale image) for visualization purposes.

can suppress details to be introduced by the temporal coupling. The infimal convolution approach allows to favor the optical flow information without over-regularizing those parts of the image, where the flow estimation seems to be faithful.

Note that – even apart from the infimal convolution – our regularization term introduces a novel type of coupling as the temporal and spatial part of the energy are in a joint ℓ^2 norm at each pixel. Previous techniques merely considered their sum.

Figure 2 demonstrates the behavior of the infimal convolution by illustrating the division of one frame into the two parts w and $z = u - w$ of (3). Areas in which the optical flow estimation is problematic are visible in the w variable and hence mostly regularized spatially. All other areas are dominated by strong temporal regularization.

2.2 Multiframe Motion Coupling

A key aspect of our approach is the temporal coupling of the (unknown) *high resolution frames* u via the warping operator W . It is based on the usual color constancy assumption, and

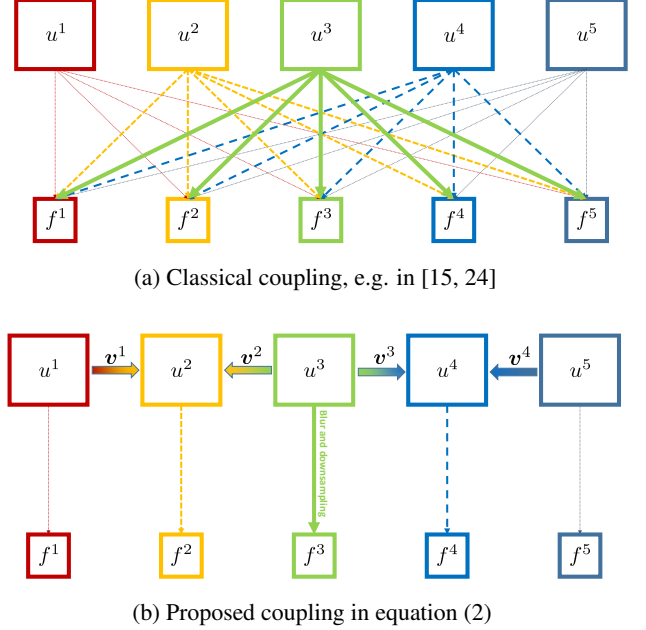


Figure 3: Illustrating different kinds of temporal couplings for an examples sequence of 5 frames: The upper image (Subfigure (a)) shows how classical methods couple the estimated high resolution frames with the input data: Every high resolution frame is linked to every low resolution frame via a warping, blur, and downsampling. The lower image (Subfigure (b)) illustrates the proposed coupling. Each frame is merely coupled to its corresponding low resolution version. The temporal information is introduced by coupling the high resolution frames by warping terms directly.

allows to couple the entire sequence in a spatio-temporal manner using only linearly many flow fields v . Figure 3 illustrates the difference of the temporal coupling of previous energy minimization techniques and the proposed method. Besides only requiring linearly many flow fields, the high resolution frames are estimated jointly such that temporal consistency is enforced directly. Note that the energies (1), or the ones of [13, 14] decouple and solve for each high resolution frame separately with the temporal conformance being introduced by the temporal consistency of the input frames f^i only. Finally, the high resolution flow fields enter the reconstruction procedure directly without being degraded by blur and downsampling operators.

Forward-Backward-Flow: During our numerical experiments we found that using a one-directional frame coupling, i.e. either estimating flows $u^i \rightarrow u^{i+1}$ or flows $u^{i+1} \rightarrow u^i$ for all i , introduces significant artifacts. We suspect these artifacts to arise from a numerical diffusion process caused by the interpolation error that is introduced by the concatenation of n (bicubic-)warping operators. In fact, a pure forward flow model $u^i \rightarrow u^{i+1}$ causes the first frame to be overly

sharp while the last frame is rather blurred and vice versa for the pure backward flow. Interestingly, our idea of alternately changing the flow direction, i.e. using $u^{i+1} \rightarrow u^i$ and $u^{i-1} \rightarrow u^i$ (see Figure 3), gave significantly better results than using two flows $u^{i+1} \rightarrow u^i$ and $u^i \rightarrow u^{i+1}$ per pair of consecutive images.

3 Optimization

The optimization is performed in a two-step procedure: We first compute the optical flow on the low resolution input frames, and upsample the flow to the desired resolution using bicubic interpolation. Then we solve the super resolution problem (2).

3.1 Optical Flow Estimation

The optical flow \mathbf{v}^i on low resolution input frames f^i is calculated via

$$\mathbf{v} = \arg \min_{\mathbf{v}^i} \sum_{i=2}^n D(\mathbf{v}_i) + \beta \sum_{j=1}^2 \|\nabla \mathbf{v}_j^i\|_{H^e},$$

$$D(\mathbf{v}) = \begin{cases} \|u^{i-1}(x) - u^i(x + \mathbf{v}^i(x))\|_1 \\ + \|\nabla u^{i-1}(x) - \nabla u^i(x + \mathbf{v}^i(x))\|_{1,1} & \text{if } i \text{ even,} \\ \|u^i(x) - u^{i-1}(x + \mathbf{v}^i(x))\|_1 \\ + \|\nabla u^i(x) - \nabla u^{i-1}(x + \mathbf{v}^i(x))\|_{1,1} & \text{if } i \text{ odd,} \end{cases} \quad (4)$$

where the term D accounts for the idea of using forward-backward-flows, and consists of two terms, one that models *brightness constancy* and one that models *gradient constancy*. Note that (4) describes a series of $n - 1$ time-independent problems. To solve each of these problems we follow well-established methods [23, 27, 29] and first linearize the brightness- and gradient constancy terms using a first order Taylor expansion with respect to the current estimate $\tilde{\mathbf{v}}^i$ of the flow field resulting in a convex energy minimization problem for each linearization. We exploit the well-known iterative coarse-to-fine approach [2, 3], which solves the problem on blurred and subsampled images first and uses an upscaled result as initial value for the next level. On each level the problem is solved several times treating the previously calculated \mathbf{v} as new a-priori solution $\tilde{\mathbf{v}}$ (see for example [27, 29]). The result \mathbf{v} is filtered once at the end of each level using a two-dimensional median filter [26]. A detailed evaluation of this strategy can be found in [23]. We use a primal-dual algorithm [18, 6] to solve the convex subproblems within the coarse-to-fine pyramid. Our implementation uses the FlexBox framework [7] which is freely available online to apply the primal-dual algorithm. It is based on MATLAB, but comes with an optional C++ and CUDA module to enhance the performance.

3.2 Super Resolution

Unlike previous approaches, the super resolution problem (2) does not simplify to a series of time-independent problems, since individual frames are correlated by the flow. Consequently, the problem has to be solved in the whole space/time domain. First, we want to deduce that (2) can be rewritten in the form

$$\arg \min_{u,w} \|\mathcal{A}u - f\|_1 + \alpha \left\| \begin{pmatrix} \nabla w \\ \kappa \mathcal{W}w \end{pmatrix} \right\|_{2,1} + \alpha \left\| \begin{pmatrix} \kappa \nabla(u - w) \\ \mathcal{W}(u - w) \end{pmatrix} \right\|_{2,1}, \quad (5)$$

where $u = (u^1, \dots, u^n)$, $f = (f^1, \dots, f^n)$, and $\mathcal{A} = \text{diag}(DB, \dots, DB)$ denotes a linear operator, i.e. a matrix in the discrete case after vectorization of the images u , that consists of the downsampling and blur operators. We chose the blur operators as Gaussian blur with variance dependent on the magnification factor, i.e. $\sigma = 1.2$ for a factor of 4, following previous methods [13, 12]. Similarly, the gradients on w and $u - w$ are block-diagonal operators consisting of the gradient operators of the single frames along the diagonal. The operator \mathcal{W} is also linear and can be seen as a motion-corrected time derivative. The notation $\|\cdot\|_{2,1}$ is used to denote the sum of the ℓ^2 norms of the vector formed by two entries from the gradient and one entry from the warping operator \mathcal{W} .

Based on the flow fields \mathbf{v} from the first step, we write the functions of the form $u^i(x + \mathbf{v}^i(x))$ as $W^i u^i$, where the W^i are the so-called interpolation or *warping* operators. Note that the operators W^i are linear operators for typical interpolation schemes such as bicubic interpolation, which means they can be expressed as a matrix to be multiplied with the vectorized image frames. Each row of the matrix W^i consists of weights according to a bicubic interpolation at a particular pixel, such that $u^i(x + \mathbf{v}^i(x)) \approx W^i u^i$. The overall matrix \mathcal{W} is the final representation for the temporal consistency term, which – due to the idea of the forward-backward-flows – is of the form

$$\mathcal{W} = \frac{1}{h} \begin{bmatrix} I & -W^1 & 0 & \dots & \dots & 0 \\ 0 & W^2 & -I & 0 & \dots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & 0 & -W^{n-2} & I & 0 \\ 0 & \dots & \dots & 0 & -I & W^{n-1} \\ 0 & \dots & \dots & 0 & 0 & 0 \end{bmatrix},$$

where h can be seen as a temporal weighting or step size parameter. Note that the last row of \mathcal{W} returns zero. In the interpretation of \mathcal{W} as a motion-corrected time derivative, the latter has the meaning of zero Neumann boundary conditions, which are the most common type of boundary conditions in image reconstructions.

Similar to the flow problem, we used an implementation of the primal-dual algorithm [18, 6] with preconditioning and adaptive stepsizes provided by the PROST framework [16] to

compute the minimizer. PROST has a MATLAB interface with the minimization running in CUDA.

For the sake of reproducibility we will make our code publicly available upon acceptance of this manuscript.

4 Numerical Results

For our implementation we choose static parameters $\alpha = 0.01$, $\beta = 0.1$ and $\kappa = 0.5$ across all experiments. The temporal step size h is estimated automatically as the ratio of warp energy to gradient energy on a bicubic estimate u_0 , taken as a vector over all frames:

$$h = \frac{\|\mathcal{W}u_0\|_1}{\|\partial_x u_0\|_1 + \|\partial_y u_0\|_1}.$$

Since the warping part is multiplied with h^{-1} , it provides an image-adaptive way to make sure that the spatial and temporal regularity terms are in the same order of magnitude.

To be able to super resolve color videos we follow the common approach from [28, 10, 11] to transform the image sequence into a YCbCr color space and super resolve the luminance channel Y with our variational method only. The chrominance channels Cr and Cb are upsampled using bicubic interpolation. Since almost all detail information is concentrated in the luminance channel, this simplification yields almost exactly the same peak signal-to-noise ratio (PSNR) as super resolving each channel separately.

4.1 Comparison to other Methods

We evaluated the presented algorithm on several scenes with very different complexity and resolution. Included in our test set is one simple synthetic scene consisting of a planar motion of the London subway map (*tube*), shown in Figure 1, the four sequences from [13] (*calendar*, *city*, *foliage*, *walk*), three sequences from [12, 20] (*foreman*, *temple*, *penguins*), and four sequences from a realistic and modern UHD video sequence (*sheets*, *wave*, *surfer*, *dog*) [22] subsampled to 720p, that contain large non-linear motions and complex scene geometries.

For all videos we created input data by using a bicubic downsampling and clipped the resulting values back to a range of $[0, 1]$. For the sake of this comparison we focused on an upsampling factor of 4, although our variational approach is able to handle arbitrary positive real upsampling factors in a straight forward fashion - a feature which would require an entirely new model and extensive retraining for neural network type of approaches.

We evaluate nearest neighbor (NN) and bicubic interpolation (Bic), Video Enhancer [19] (a commercial upsampling software), the variational approach [14] (MFSR), as well as the learning based techniques Deep Draft [12], VSRnet [10], and VDSR [11], along with our proposed method for 13 frames of the *tube*, *city*, *calendar*, *foliage*, *walk* and *foreman* sets and 5 frames of the larger *temple*, *penguins*, *sheets*, *surfer*,

wave and *dog* sets. The PSNR and structural similarity index measure (SSIM) [25] were determined for the central image of each sequence after cropping 20 pixels at each boundary.

SSIM	NN	Bic	[19]	[14]	MMC	[12]	[10]	[11]
<i>tube</i>	0.799	0.852	0.898	0.877	0.942	0.881	0.901	0.918
<i>city</i>	0.596	0.634	0.702	0.653	0.760	0.725	0.68	0.688
<i>calendar</i>	0.621	0.659	0.706	0.686	0.762	0.739	0.705	0.726
<i>foliage</i>	0.760	0.804	0.809	0.809	0.870	0.853	0.831	0.836
<i>walk</i>	0.776	0.840	0.858	0.825	0.893	0.841	0.875	0.886
<i>foreman</i>	0.88	0.920	0.924	0.923	0.949	0.928	0.941	0.953
<i>temple</i>	0.835	0.88	0.893	0.878	0.922	0.873	0.916	0.927
<i>penguins</i>	0.939	0.964	0.966	0.965	0.970	0.951	0.976	0.979
<i>sheets</i>	0.948	0.973	0.978	0.972	0.981	0.974	0.979	0.979
<i>surfer</i>	0.967	0.981	0.979	0.945	0.984	0.963	0.985	0.986
<i>wave</i>	0.941	0.958	0.963	0.955	0.971	0.963	0.964	0.966
<i>dog</i>	0.955	0.972	0.974	0.970	0.976	0.969	0.977	0.977
PSNR								
<i>tube</i>	18.63	20.32	21.73	20.82	23.73	21.39	21.88	22.36
<i>city</i>	23.35	23.95	24.75	24.23	25.52	24.90	24.45	24.60
<i>calendar</i>	18.07	18.84	19.49	19.20	20.38	19.96	19.36	19.63
<i>foliage</i>	21.21	22.41	23.19	22.40	24.05	23.47	23.00	23.16
<i>walk</i>	22.74	24.69	25.37	23.98	26.75	24.98	25.95	26.40
<i>foreman</i>	26.40	28.95	29.31	28.39	31.71	29.52	31.01	32.54
<i>temple</i>	24.15	25.76	26.29	25.84	27.49	25.69	27.39	27.90
<i>penguins</i>	29.17	32.27	32.82	32.54	33.20	30.64	34.63	35.00
<i>sheets</i>	29.68	32.76	33.73	32.27	34.20	33.01	33.86	33.95
<i>surfer</i>	30.59	33.29	33.29	29.11	34.15	32.45	34.42	34.96
<i>wave</i>	30.73	32.20	32.82	31.85	33.80	32.45	33.03	33.33
<i>dog</i>	32.58	34.81	35.07	34.15	35.47	34.08	35.63	35.71

Table 1: SSIM and PSNR values (4x upsampling) from left to right: nearest neighbor, bicubic, commercial VideoEnhancer software [19], Multi-Frame Super resolution [14], MMC (our approach), DeepDraft ensemble learning [12], VSRnet deep convolutional network (with adaptive motion compensation) [10], VDSR (image upsampling using very deep convolutional networks) [11].

The results for all test sequences are shown in Table 1. We structured the methods into three categories; simple interpolation based methods, variational super resolution approaches that utilize temporal information but do not require any training data, and deep learning methods. We indicate the three categories by vertical lines in the tables.

As we can see, our method consistently improves upon simple interpolation techniques and also clearly outperforms the competing variational approaches. In comparison to the learning based approaches, our explicit model based technique seems to be superior on those sequences that contain small motion or a high frame rate. On sequences with particularly large motion and strong occlusions, the very deep convolutional neural network [11] performs very well, possibly because it does not rely on any motion information but produces high quality results frame by frame purely based on learned information.

Besides the fact that our approach remains competitive even for the aforementioned challenging data sets in terms

of the PSNR values, we illustrate in several videos in the supplementary material that the PSNR value of a single frame does not quantify the visual quality of the video produced by each approach well: The temporal consistency of successive frames is extremely important for a human observer to perceive the super resolved video as visually pleasing. A lack of temporal consistency immediately yields a disturbing flickering effect, which is not present in the results produced by the proposed technique.

For a visual inspection of single frames, we present the super resolution results obtained by various different methods on a selection of four different data sets in Figure 4. The first column, the *calendar* data set, contains slow and steady motions. We can see that the variational approach MFSR with standard parameters cannot compete with our approach in sharpness, whereas the learning based techniques VSRnet and VDSR are sharp but have difficulties making the writing "MAREE" readable again. The following three columns contain scenes with fast motions and occlusions. As we can see, the strong spatial regularization of the MFSR suppresses fine details for these scenes, yet lowering this regularization would however lead to severe motion artifacts. Our technique relies on motion information as much as possible, but only when appropriate and with that produces results competitive to learning based techniques that require large amounts of training data.

The sequence *penguins* is the only one where fast and complex motion caused visual problems as illustrated in the zoom of the inset figure. While the PSNR value still improved by 1dB over the bicubic interpolation some small artifacts are visible along the penguins feet and wings. Note that increasing the parameter κ enforces a stricter spatial regularity baseline and can avoid such artifacts in scenes with quick motion, at the cost of temporal details. For example, we can increase the PSNR value on the *tube* sequence to 24.27 dB by choosing $\beta = 0.5$ and setting $\kappa = 0.1$. It is important to note that the results in Table 1 were obtained with static parameters that we found to yield a good and robust trade off. Future work could include the application of techniques like [1] that provide an automated parameter optimization algorithm for κ given a representative sample of expected video data.



4.2 Run Times

We report run times of 5 minutes on average for jointly super resolving 5 frames to a resolution of 1280x720 pixels, using a laptop with a NVIDIA GTX 1080. Although these runtimes are on a modern GPU, the optical flow problem and the super resolution problem are both implemented in general-purpose primal-dual toolboxes that are neither optimized for the par-

ticular problem we are solving, nor do they communicate directly. Further increases in speed can therefore be obtained by porting the entire application to a specialized GPU framework.

4.3 Further Considerations and Extensions

In this paper, we focused on presenting the super resolution framework for optimizing for the high resolution frames u . Our technique can easily be extended by optimizing for further parameters, e.g. by additionally estimating for motion blur kernels b as done in [13, 14]. Additionally, one can incorporate alternating minimization steps between our energy (2), and estimating the high resolution flow fields. Our numerical experiments with both of the aforementioned techniques changed our results in the second digit of the PSNR values only. The latter does not justify the additional computational expenses for the test sequences we used, but can pose an interesting option for videos containing significant (and temporally changing) motion blur.

5 Conclusions

We have proposed a variational super resolution technique based on a multiframe motion coupling of the unknown high resolution frames. The latter enforces temporal consistency of the super resolved video directly and requires only as many optical flow estimation as there are frames. By combining spatial regularity and temporal information with an infimal convolution instead of a sum, our method adapts the strength of spatial and temporal smoothing automatically. We provided an extensive numerical comparison which demonstrates that the proposed method outperforms interpolation approaches as well as a competing state-of-the-art variational super resolution method. Our approach is competitive with learning based approaches which have to rely on large amounts of training data. For small motions or sufficiently high frame rate, our results are temporally consistent and avoid flickering effects visible in deep learning approaches. Failure cases of our method are observed in areas of very quick motions, where the optical flow does not yield any reasonable inter-frame connections.

References

- [1] M. Benning, C.-B. Schönlieb, T. Valkonen, and V. Vlačić. Explorations on anisotropic regularisation of dynamic inverse problems by bilevel optimisation. *arXiv preprint arXiv:1602.01278*, 2016.
- [2] M. J. Black and P. Anandan. The robust estimation of multiple motions: Parametric and piecewise-smooth flow fields. *Computer vision and image understanding*, 63(1):75–104, 1996.
- [3] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert. High accuracy optical flow estimation based on a theory for warping. In *European conference on computer vision*, pages 25–36. Springer, 2004.
- [4] M. Burger, H. Dirks, and C.-B. Schönlieb. A variational model for joint motion estimation and image reconstruction. *arXiv preprint arXiv:1607.03255*, 2016.
- [5] A. Chambolle and P.-L. Lions. Image recovery via total variation minimization and related problems. *Numerische Mathematik*, 76(2):167–188, 1997.
- [6] A. Chambolle and T. Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of Mathematical Imaging and Vision*, 40(1):120–145, 2011.
- [7] H. Dirks. A flexible primal-dual toolbox. *arXiv preprint*, 2016. <http://www.flexbox.im>.
- [8] S. Farsiu, M. D. Robinson, M. Elad, and P. Milanfar. Fast and robust multiframe super resolution. *IEEE transactions on image processing*, 13(10):1327–1344, 2004.
- [9] M. Holler and K. Kunisch. On infimal convolution of tv-type functionals and applications to video and image reconstruction. *SIAM Journal on Imaging Sciences*, 7(4):2258–2300, 2014.
- [10] A. Kappeler, S. Yoo, Q. Dai, and A. K. Katsaggelos. Video super-resolution with convolutional neural networks. *IEEE Transactions on Computational Imaging*, 2(2):109–122, 2016.
- [11] J. Kim, J. Kwon Lee, and K. Mu Lee. Accurate image super-resolution using very deep convolutional networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR Oral)*, June 2016.
- [12] R. Liao, X. Tao, R. Li, Z. Ma, and J. Jia. Video super-resolution via deep draft-ensemble learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 531–539, 2015.
- [13] C. Liu and D. Sun. On bayesian adaptive video super resolution. *IEEE transactions on pattern analysis and machine intelligence*, 36(2):346–360, 2014.
- [14] Z. Ma, R. Liao, X. Tao, L. Xu, J. Jia, and E. Wu. Handling motion blur in multi-frame super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5224–5232, 2015.
- [15] D. Mitzel, T. Pock, T. Schoenemann, and D. Cremers. Video super resolution using duality based TV-L1 optical flow. In *Pattern Recognition*, pages 432–441. Springer, 2009.
- [16] T. Möllenhoff, E. Laude, M. Moeller, J. Lellmann, and D. Cremers. Sublabel-accurate relaxation of nonconvex energies. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. <https://github.com/tum-vision/prost>.
- [17] K. Nasrollahi and T. B. Moeslund. Super-resolution: a comprehensive survey. *Machine vision and applications*, 25(6):1423–1468, 2014.
- [18] T. Pock, D. Cremers, H. Bischof, and A. Chambolle. An algorithm for minimizing the Mumford-Shah functional. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 1133–1140. IEEE, 2009.
- [19] Infognition Co. Ltd. Videoenhancer 2 software, version 2.1.
- [20] Xiph.org, redistributable Video Test Media Collection. <https://media.xiph.org/video/derf/>.
- [21] E. Shechtman, Y. Caspi, and M. Irani. Space-time super-resolution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(4):531–545, 2005.
- [22] Sony Corporation. Sony 4k uhd surfing screen test demo. CC-BY License.
- [23] D. Sun, S. Roth, and M. J. Black. A quantitative analysis of current practices in optical flow estimation and the principles behind them. *International Journal of Computer Vision*, 106(2):115–137, 2014.
- [24] M. Unger, T. Pock, M. Werlberger, and H. Bischof. A convex approach for variational super-resolution. In *Pattern Recognition*, pages 313–322. Springer, 2010.
- [25] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *Image Processing, IEEE Transactions on*, 13(4):600–612, 2004.
- [26] A. Wedel, D. Cremers, T. Pock, and H. Bischof. Structure-and motion-adaptive regularization for high accuracy optic flow. In *ICCV*, pages 1663–1668, 2009.
- [27] A. Wedel, T. Pock, C. Zach, H. Bischof, and D. Cremers. An improved algorithm for TV-L1 optical flow. In *Statistical and Geometrical Approaches to Visual Motion Analysis*, pages 23–45. Springer, 2009.
- [28] J. Yang, J. Wright, T. S. Huang, and Y. Ma. Image super-resolution via sparse representation. *IEEE transactions on image processing*, 19(11):2861–2873, 2010.
- [29] C. Zach, T. Pock, and H. Bischof. A duality based approach for realtime TV-L1 optical flow. In *Pattern Recognition*, pages 214–223. Springer, 2007.
- [30] Z. Zhang and V. Sze. Fast: Free adaptive super-resolution via transfer for compressed videos. Available on ArXiv, <https://arxiv.org/abs/1603.08968>, 2016.



calendar dataset, ground truth zoom



walk dataset, ground truth zoom



foreman dataset, ground truth zoom



wave dataset, ground truth zoom



Nearest, PSNR 18.07



Nearest, PSNR 22.74



Nearest, PSNR 26.40



Nearest, PSNR 30.73



MFSR [14], PSNR 19.20



MFSR, PSNR 23.98



MFSR, PSNR 28.39



MFSR, PSNR 31.85



MMC (proposed method), PSNR 20.38



MMC, PSNR 26.75



MMC, PSNR 31.71



MMC, 33.80



VSRnet [10], PSNR 19.36



VSRnet, PSNR 25.95



VSRnet, PSNR 31.01



VSRnet, PSNR 33.03



VDSR [11], PSNR 19.63



VDSR, PSNR 26.40



VDSR, PSNR 32.54



VDSR, PSNR 33.33

Figure 4: Super resolution by a factor of 4, zoom into datasets *calendar*, *walk*, *foreman*, *wave*. PSNR values computed as described in Section 5. One can see the effective resolution increase of our method for the writing in *calendar*, faces in *walk* and wave front in *wave* as well as the robustness of the approach for the challenging *foreman* sequence.